(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

CORRECTED VERSION

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 March 2004 (04.03.2004)

PCT

(10) International Publication Number
WO 2004/019230 A3

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/US2003/026025

(22) International Filing Date: 20 August 2003 (20.08.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/404,581 20 August 2002 (20.08.2002) US
10/293,859 13 November 2002 (13.11.2002) US

(71) Applicant *(for all designated States except US)*: **MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.** [JP/JP]; Matsushita IMP Bldg., 19F, 1-3-7, Shiromi, Shuo-ku, Osaka 540-6319 (JP).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **GUO, Jinhong, Katherine** [US/US]; 6 Tiffany Court, West Windsor, NJ 08550 (US). **MA, Yue** [US/US]; 6 Tiffany Court, West Windsor, NJ 08550 (US).

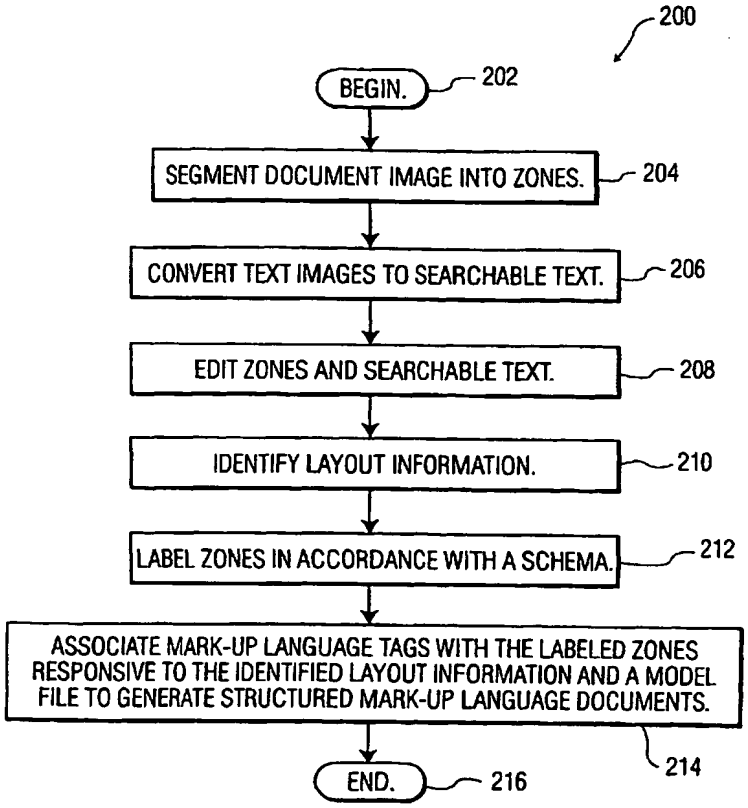
(74) Agent: **NIGON, Kenneth, N.**; RatnerPrestia, P.O. Box 980, Valley Forge, PA 19482 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE,

[Continued on next page]

(54) Title: METHOD, SYSTEM, AND APPARATUS FOR GENERATING STRUCTURED DOCUMENT FILES



(57) Abstract: A method, system, apparatus, and graphical user interface (GUI) for generating structured document files from a document image is disclosed. Structured document files are generated by segmenting the document image into one or more zones containing respective text images, converting the respective text images to digital text, automatically identifying layout information for each of the one or more zones, labeling each of the one or more zones in accordance with a schema, and automatically associating mark-up language tags with the labeled zones to generate the structured document files responsive to the identified layout information and a model file.

WO 2004/019230 A3

ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declaration under Rule 4.17:

— *of inventorship (Rule 4.17(iv)) for US only*

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
25 March 2004

(48) Date of publication of this corrected version:
29 April 2004

(15) Information about Correction:
see PCT Gazette No. 18/2004 of 29 April 2004, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.